如何让博物馆更具互动性?
案例研究艺术聊天机器人

菲利普 · J · 库查 [1,2], 巴托什 · 格拉贝克 [1], 西蒙 · D · 特罗奇米亚克 [1], 安娜 · 沃罗布莱夫斯卡 [1]
filip.kucia@gmail.com
{filip.kucia.stud,bartosz.grabek.stud,szymon.trochimiak.stud,anna.wroblewska1}@pw.edu.pl
[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[2]Samsung Research and Development Institute Poland, Warsaw, Poland

ABSTRACT

由大型语言模型（LLMs）驱动的对话代理在教育环境中越来越被使用，特别是在个人封闭数字环境中的应用，然而它们在物理学习环境如文化遗产地、博物馆和艺术画廊中的潜在采用仍相对未被探索。

在这项研究中，我们介绍了艺术型聊天机器人[1]，这是一个基于语音到语音 RAG（检索增强生成）的聊天系统，旨在支持非正式学习并提升参观者在庆祝波兰华沙美术学院媒体艺术系成立 15 周年的现场艺术展览期间的参与度。

问答（QA）聊天机器人使用由组织者提供的 226 份文件中整理出的特定领域知识库中的内容来回应以波兰语进行的自由形式口头提问，这些文件包括教师信息、艺术杂志、书籍和期刊等。

我们描述了该系统的架构和用户交互设计的关键方面，并讨论了在公共文化场所部署聊天机器人的实际挑战。

我们的研究结果基于互动分析，表明诸如艺术型聊天机器人之类的聊天机器人能够有效维持回应与展览内容相关（60%的回复直接相关），即使面对超出目标领域的不可预测查询时也是如此，这显示了它们在增加公众文化场所交互性方面的潜力。

在演示展示期间，观众将被邀请查询我们的艺术型聊天机器人，它采用了一个人工艺术策展人的身份，这一角色包括回应问题的同时评估这些问题与展览的相关性。演示视频的链接可从这里获得。

***K*eywords** 自然语言处理 · 语音接口 · 人机交互 · 大型语言模型

1 介绍

人工智能（AI）及其具体的自然语言处理（NLP）领域的快速发展，导致大型语言模型（LLMs）的广泛应用 [1]。虽然 LLMs 可以在各种环境中使用，但它们特别适合于开发对话代理，也称为聊天机器人和语音助手 [2, 3]。这些系统由最先进的模型如 GPT 驱动，在教育 [4] 领域已被证明非常有效，是智能辅导系统和个人

[1]https://github.com/cinekucia/artistic-chatbot-cikm2025

学习平台发展的基础。通过利用其生成有意义的人类般响应的能力，这些助手大大提升了用户互动，并提供了更加个性化和基于语境的对话 [5]。

大多数基于 LLM 的聊天机器人的教育应用集中在个性化学习场景上 [6, 7]，也就是说，在孤立的数字环境中使用，并通过文本界面进行交互。然而，相对较少的研究探讨了对话代理在更具社交性、共享性和互动性的环境中的潜力，例如教室或文化及公共教育场所，包括博物馆和画廊的展览场地 [8, 9]。

若干先前的研究强调了在文化遗产和博物馆场地中使用聊天机器人的有效性，以增强游客的参与度并更方便地获取信息 [10, 11]. 一个部署在意大利庞贝考古遗址的聊天机器人系统利用语义分析从书面用户查询中提取主题，并结合第三方网络服务（例如维基百科、TripAdvisor）和静态知识库中的信息，为游客的问题制定答案 [12, 13]. 另一项研究则采用多模态大型语言模型（MLLMs）来解决情境视觉问答问题 [14]. 该系统将视觉输入与来自维基百科内容数据库的相关艺术品描述相结合，确保提供科学准确的信息。

在这项工作中，我们展示了艺术型聊天机器人，这是一个语音到语音的问答（QA）聊天机器人，为纪念华沙美术学院媒体艺术系 15 周年庆典而开发并部署。该系统允许参观者就学院、艺术家、展览和艺术品提出自由形式的口语问题，并收到基于主办方事先提供的领域特定语料库的人类般口语回答。

2 系统设计与实现

聊天机器人的开发涉及两个阶段：一个数据预处理阶段以构建专门的知识库，以及一个设计用于处理用户交互的推理管道。

2.1 数据预处理

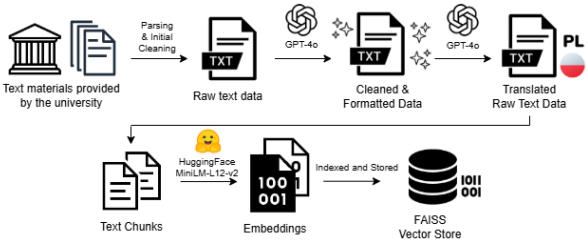

图 1: 数据预处理管道

初始阶段的重点是从华沙美术学院提供的 226（159 份原始 PDF 文档和 67 份艺术家及学院员工的传记）中准备知识库（参见图 1）。这些源文件存在相当大的挑战，主要是由于复杂的布局、多种语言的存在——如波兰语、普通话、英语、德语和法语——以及非结构化数据中常见的内在不一致性。虽然约 92%的数据是波兰语的，但许多文档包含其他语言的片段，通常在同一页面上，这增加了整体解释和处理的复杂性。为了将这些原始数据精炼成高质量且一致的资源，提取的文本文件被清理并使用多语言大型语言模型——GPT-4o [15] 翻译成了波兰语。

数据集文档被分割成 11,596 个较小的、重叠的部分，以促进有效的检索增强生成（RAG）[16]。这些部分限制为 5,000 个字符，并且有 200 个字符的重叠，以确保各段落之间的上下文连续性。平均而言，每个文档被分割成大约 51 个块（中位数：19.5），尽管数量因长度和结构的不同而有所变化（每份文档从 1 到 650 块不等）。然后，每个生成的文本块使用 *sentence-transformers*/语义相似度多语言 *MiniLM-L12-v2* 模型被转换为密集向量表示形式（嵌入）。随后，这些嵌入被索引到 FAISS（Facebook AI 相似性搜索）[17] 向量存储中，以便快速查找最相关的块。

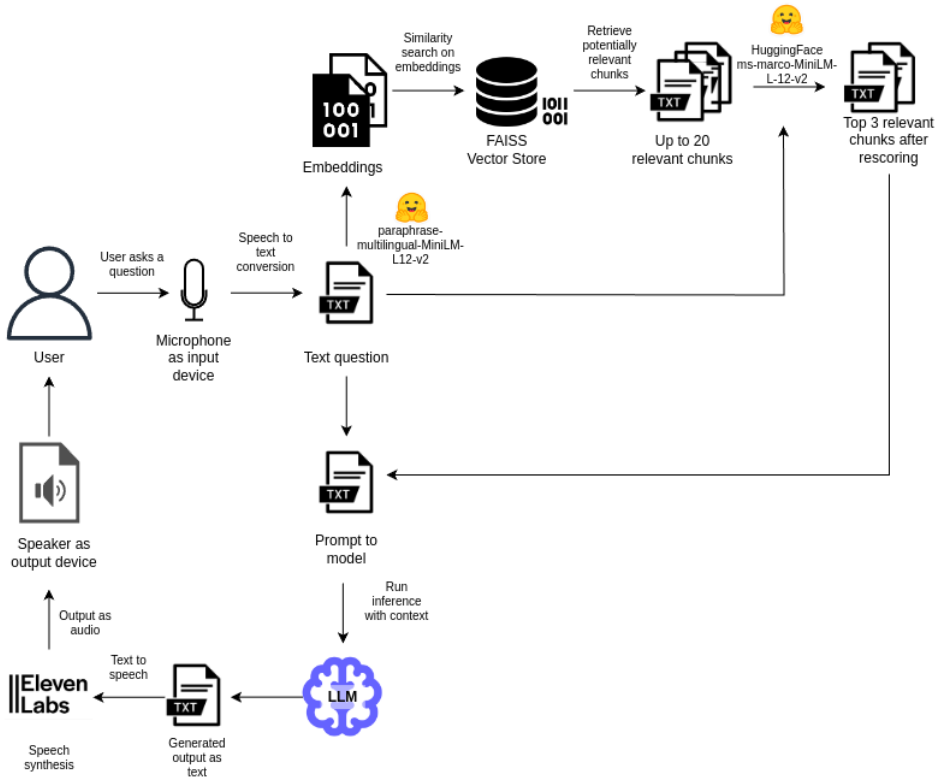

图 2: 推理管道

2.2 推理管道

第二阶段，推理管道，是用户交互流程的关键（见图 2）。用户参与从语音输入开始。系统会主动监听预定义的触发词（参见第 2.3 节），然后捕获用户的查询。这使我们能够避免不必要的触发，并明确表明查询的开始。在成功将查询转录为文本后，调用了 RAG 机制。这涉及一个两步检索过程。首先，嵌入了用户查询。FAISS 向量存储随后执行了一个初步相似性搜索，从索引的知识库中检索出前 20 个潜在相关片段。此初始检索优先考虑速度而非精细的相关性；因此，我们还需要一个后续的重新排序步骤。它使用预训练的 CrossEncoder 模型交叉编码器*/ms-marco-MiniLM-L12-v2*，更精确地评估每个 20 检索到的段落与原始查询的相关性。仅选择由重新排名者确定的得分最高的 3 个片段作为最终响应生成的上下文基础。然后将选定的上下文和用户的查询一起用于构建 LLM 的提示。响应生成由 GPT-4o-mini 模型处理。提示包括一个动态选择的系统消息，在会话开始时随机选择响应风格，包括“普通”、“学术”或“随意”的模板。此系统提示定义了聊天机器人的个性以区分系统的回答风格。提示还提供了关于展览的重要上下文信息，如其位置、展览时间，并概述了具体的对话指南，包括仅用波兰语回应和避免直接引用检索到的上下文。最后，文本响应使用 ElevenLabs 文本转语音 API 合成回可听见的声音。每个完整的交互，包括时间戳、用户输入查询、所使用的系统提示风格以及最终生成的响应，都被系统地记录下来，以实现对聊天机器人性能的持续监控和后续分析。

2.3 用户交互

展览的物理设置包括一个安装在天花板上的麦克风，位于房间中央，以及四个安装在角落里的扬声器。外设连接到附近的一台运行聊天机器人的 PC 工作站，并实时记录用户响应（参见图 3）。

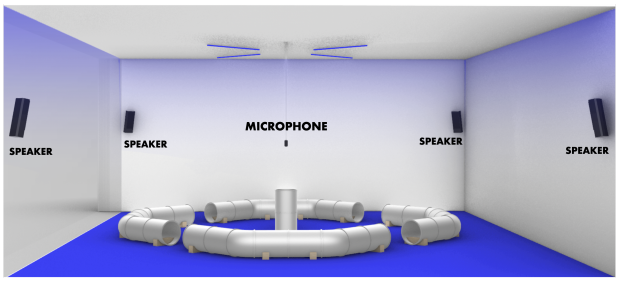

图 3: 聊天机器人物理设置

为了与艺术聊天机器人进行互动，访客需要走到悬挂的麦克风前并说出四个指定的触发表达式之一（“你好”，“欢迎”，“问题”或“I have a question”[2]）。系统在识别到触发词后会向访客打招呼，从而表示它已准备好接收问题。这一机制确保了输入是故意为之，将用户互动与背景噪音区分开来。

主要目标是鼓励访客提问与展览、参展艺术家或媒体艺术学院历史相关的问题。一旦问候过后，用户可以在该主题范围内几乎提出任何问题。聊天机器人设计为单轮交互，意味着它一次只回答一个问题，并且不会保留之前交流的上下文信息。在接受新问题之前，聊天机器人必须完成对当前问题的回答生成并以音频形式播放完毕。访客被期望等待语音回答完全结束后再开始另一个查询。这一限制确保了清晰度，防止了打断，并保持了交互的一致性。

每个问答循环结束后，系统会自动返回初始状态，监听触发短语以重新激活聊天机器人。这种交互模型通过语音输入与人工智能艺术策展人进行模拟对话，为用户提供了一种直接、易于访问的方式，来接触展览的内容和概念。

3 交互分析

在展览对公众开放的那个月里，总共提出了 727 个问题。我们使用 LLM 作为评判的技术评估了这些问题和大型语言模型的回答（遵循 [18]）。我们提示了一个大型语言模型去检查问题的完整性、扩展问题、评价回答的相关性并分类问题。主要挑战之一是问题的完整性（参见表 1）。所有捕获的用户查询中，近三分之一的问题不完整或被提前切断了。这一问题主要是由于系统依赖于基于沉默的句尾检测所致。然而，分析表明纠正这些问题通常只需要进行很小的修改——平均大约 3.4 个字符（在 1 到 6 的范围内，以莱文斯坦距离衡量）。

表 1: 交互统计：问题（Q.）的完整性、问题的相关性以及回复（R.）与展览领域的相关性。

度量	是 (#)	没有 (#)	是 (%)
Q. completeness	497	230	68.37
Q. relevance	142	585	19.53
R. relevance	440	287	60.52

表 2: 问题-响应相关性评分分布（平均值：2.66）。注意评分范围：1（不相关）到 5（非常相关）

得分	1	2	3	4	5
Count	286	37	169	110	125
Percentage (%)	39.4	5.1	23.2	15.1	17.2

[2]来自波兰语：“你好“, “欢迎“, “问题“, “我有一个问题“

正如其他展览设置中一样，例如在 [19] 中，大多数问题——大约 600 个——被归类为简单的事实性问题，其次是大约 150 个随意和确认性问题，以及 24 个假设性问题（例如，“如果 X 是真的会发生什么”）。这种行为模式大大减少了与系统的有意义互动，并被认为是需要改进的关键领域。我们的参观者经常偏离预期的问题范围，这是公共、基于语音的装置中常见的挑战。事实上，只有五分之一的问题完全与目标领域相关（参见表 1）。然而，由于系统始终向大语言模型提供了从知识库检索到的相关展览背景，许多答案仍然集中在展览上，即使用户的问题无关（参见表 2）。

结果表明，该系统在将回应与目标语境相联系方面表现出色，这对于展览尤为重要。然而，我们发现需要进一步发展和测量输入处理及回应的相关性，特别是在开放的公共设置中，预期用户陈述会有较高的变异性。

4 限制与伦理考量

艺术聊天机器人担任了实时原型的角色，并且自然涉及了一些实用的捷径，导致了几处限制。

首先，用于 RAG 的数据集相对较小。扩大数据集是一个解决方案，但这可能会影响检索延迟，从而影响整个系统的效率。然而，这些权衡被认为是 RAG 模型推理 [20] 的不可避免和典型特征。这也是为什么在类似设置中设计可扩展聊天机器人的重点应放在通过最佳分块 [21] 和适当选择 RAG 系统参数（例如检索的块数 [22]）来有效利用上下文的原因。此外，将源材料翻译成波兰语可能会引入不准确性，主要是由于领域术语和文化背景的变化。由于艺术品收藏通常涵盖多种语言，可以使用除完全基于翻译的 RAG 流水线之外的替代方法，例如跨语言 RAG[23]，其中检索在原始语言中进行，并且仅在响应生成之前翻译最相关的块。

其次，观察到了一些用户体验（UX）问题。系统依赖检测停顿来确定用户查询的结束，这可能导致过早终止用户输入检索，特别是在嘈杂环境中或当用户说话时犹豫不决的情况下。将交互方式从语音命令激活更改为物理按钮激活以使系统听取用户声音是一种简单的问题解决方案。另一种方法，在不改变物理交互界面的情况下，可以通过增强现有的静默检测并采用更为复杂的语句结束（EoU）[24] 验证机制来实现。

大型语言模型驱动的聊天系统在公共场所，尤其是博物馆和文化遗产地中的部署需要解决潜在的伦理问题。主要关注点之一是大型语言模型和 RAG 系统的准确性。大型语言模型会产生幻觉，经常给出看似合理但实际上是错误或误导性的信息。尽管领域特定的 RAG 减少了产生幻觉的概率，但由于来源文档不准确或大型语言模型对上下文 [25] 的误解，仍然可能出现虚假信息。为了防止幻觉并提高聊天机器人的可靠性，可以考虑微调模型以承认不确定性，即在必要时用“我不知道”来回应 [26]。另一个需要考虑的伦理方面是聊天机器人回复的安全性。除非采用内容过滤和监管技术，否则系统被用户滥用或误用（例如通过提示工程）时，生成有害或不适当的内容的风险会增加。

5 结论

在公共艺术展览中部署艺术型聊天机器人展示了基于 LLM 的语音代理增强文化遗产场所互动性和参观者参与度的潜力。聊天机器人作为展览的一部分成功运行了一个月。在这为期一个月的活动中，记录了超过 727 次查询，其中有多达 60% 的回答被认为是与展览相关的，尽管只有大约 20% 的用户问题是严格符合主题的。这表明系统能够在面对不可预测且超出目标领域输入的情况下，仍保持回应紧扣展览内容的能力。然而，基于检索的回答也带来了权衡，例如处理模糊或含糊查询时灵活性受限和答案多样性减少的问题。诸如人机交互设计、简单的句尾检测以及整体回答相关性一般（平均分：2.66）等挑战被确定为未来改进的关键领域。

致谢

我们想感谢 ElevenLabs 的合作伙伴提供的文字转语音服务以及美术学院员工的合作。

参考文献

[1] Michael McTear and Marina Ashurkina. *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Apress, Berkeley, CA, 2024.

[2] Simone Gallo, Fabio Paternò, and Alessio Malizia. A conversational agent for creating automations exploiting large language models. *Personal and Ubiquitous Computing*, 28(6):931–946, December 2024.

[3] Mina Foosherian, Hendrik Purwins, Purna Rathnayake, Touhidul Alam, Rui Teimao, and Klaus-Dieter Thoben. Enhancing Pipeline-Based Conversational Agents with Large Language Models, September 2023. arXiv:2309.03748 [cs].

[4] Lasha Labadze, Maya Grigolia, and Lela Machaidze. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(1):56, October 2023.

[5] Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding Chit-Chat to Enhance Task-Oriented Dialogues, May 2021. arXiv:2010.12757 [cs].

[6] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education, August 2023. arXiv:2308.02773 [cs].

[7] Xinyu Jessica Wang, Christine Lee, and Bilge Mutlu. LearnMate: Enhancing Online Education with LLM-Powered Personalized Learning Plans and Support, March 2025. arXiv:2503.13340 [cs].

[8] Yeo-Gyeong Noh and Jin-Hyuk Hong. Designing reenacted chatbots to enhance museum experience. *Applied Sciences*, 11(16), 2021.

[9] Mario Casillo, Fabio Clarizia, Giuseppe D'Aniello, Massimo De Santo, Marco Lombardi, and Domenico Santaniello. Chat-bot: A cultural heritage aware teller-bot for supporting touristic experiences. *Pattern Recognition Letters*, 131:234–243, 2020.

[10] Giancarlo Sperlí. A deep learning based chatbot for cultural heritage. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 935 – 937, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Konstantinos Tsitseklis, Georgia Stavropoulou, Anastasios Zafeiropoulos, Athina Thanou, and Symeon Papavassiliou. Recbot: Virtual museum navigation through a chatbot assistant and personalized recommendations. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 388 – 396, New York, NY, USA, 2023. Association for Computing Machinery.

[12] M. Lombardi, F. Pascale, and D. Santaniello. An application for cultural heritage using a chatbot. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–5, 2019.

[13] Giancarlo Sperlí. A cultural heritage framework using a deep learning based chatbot for supporting tourist journey. *Expert Systems with Applications*, 183:115277, 2021.

[14] Pavan Kartheek Rachabatuni, Filippo Principi, Paolo Mazzanti, and Marco Bertini. Context-aware chatbot using MLLMs for Cultural Heritage. In *Proceedings of the 15th ACM Multimedia Systems Conference*, MMSys '24, pages 459–463, New York, NY, USA, 2024. Association for Computing Machinery.

[15] OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2025. Accessed April 12, 2025.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[17] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library, February 2025. arXiv:2401.08281 [cs].

[18] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[19] Zhaozhen Xu, Amelia Howarth, Nicole Briggs, and Nello Cristianini. Understanding visitors' curiosity in a science centre with deep question processing network. *International Journal of Artificial Intelligence in Education*, 34(3):1072 – 1101, September 2024.

[20] Michael Shen, Muhammad Umar, Kiwan Maeng, G. Edward Suh, and Udit Gupta. Towards understanding systems trade-offs in retrieval-augmented generation model inference, 2024.

[21] Kush Juvekar and Anupam Purwar. Introducing a new hyper-parameter for rag: Context window utilization, 2024.

[22] Juraj Vladika and Florian Matthes. On the influence of context size and model choice in retrieval-augmented generation systems, 2025.

[23] Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task, 2025.

[24] Oswald Zink, Yosuke Higuchi, Carlos Mullov, Alexander Waibel, and Tetsunori Kobayashi. Predictive Speech Recognition and End-of-Utterance Detection Towards Spoken Dialog Systems, September 2024. arXiv:2409.19990 [eess].

[25] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models, 2024.

[26] Xinxi Chen, Li Wang, Wei Wu, Qi Tang, and Yiyao Liu. Honest ai: Fine-tuning "small" language models to say "i don't know", and reducing hallucination in rag, 2024.